MST121 Chapter D2

The Open University

A first level interdisciplinary course

Using Mathematics

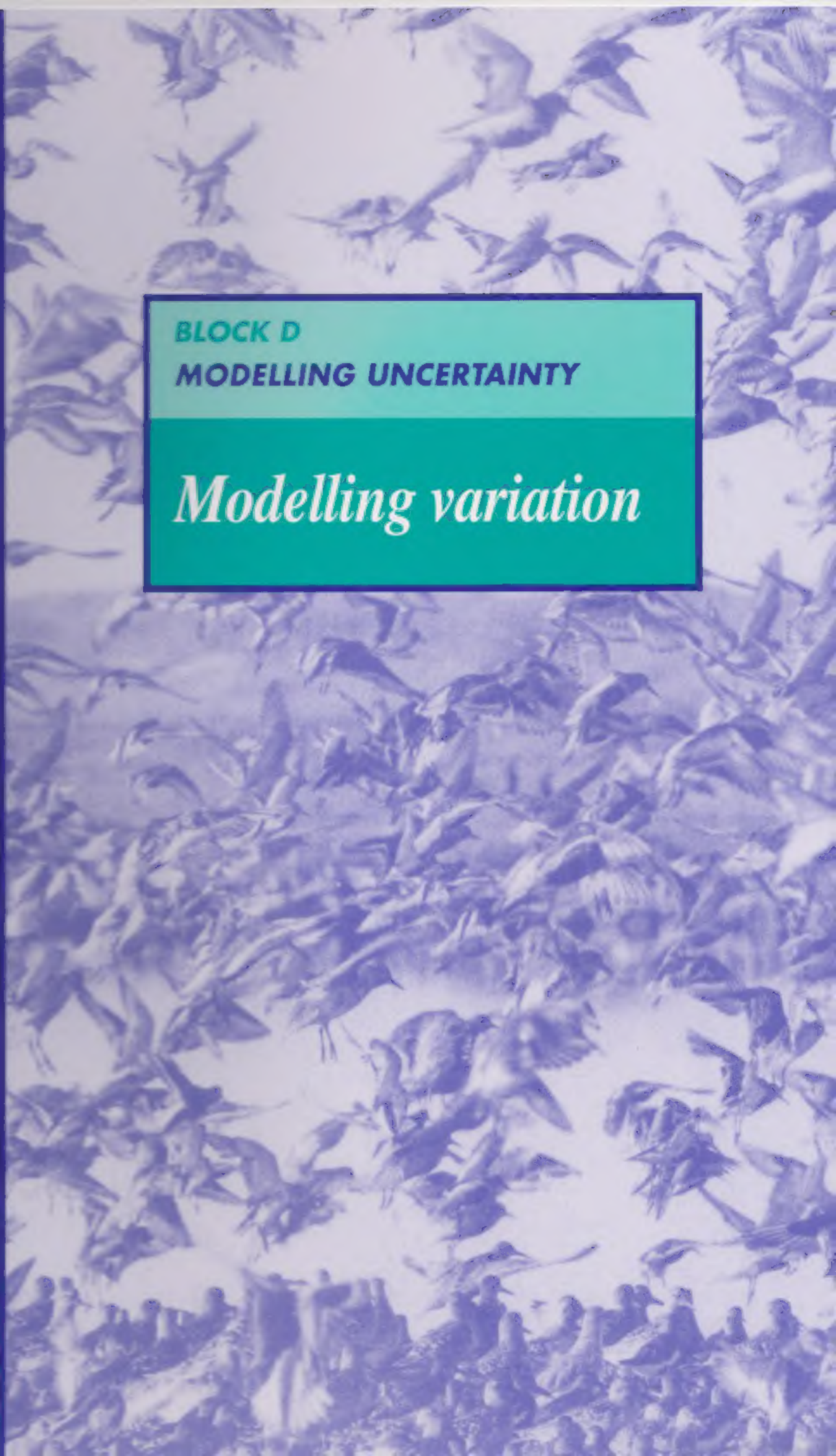
BLOCK D

MODELLING UNCERTAINTY

Modelling variation

CHAPTER

D2



The Open University

A first level interdisciplinary course

Using Mathematics

CHAPTER D2

BLOCK D

MODELLING UNCERTAINTY

Modelling variation

Prepared by the course team

About this course

This course, MST121 *Using Mathematics*, and the courses MU120 *Open Mathematics* and MS221 *Exploring Mathematics* provide a flexible means of entry to university-level mathematics. Further details may be obtained from the address below.

MST121 uses the software program Mathcad (MathSoft, Inc.) and other software to investigate mathematical and statistical concepts and as a tool in problem solving. This software is provided as part of the course.

This publication forms part of an Open University course. Details of this and other Open University courses can be obtained from the Course Information and Advice Centre, PO Box 724, The Open University, Milton Keynes, MK7 6ZS, United Kingdom: tel. +44 (0)1908 653231, e-mail general-enquiries@open.ac.uk

Alternatively, you may visit the Open University website at http://www.open.ac.uk where you can learn more about the wide range of courses and packs offered at all levels by The Open University.

To purchase a selection of Open University course materials, visit the webshop at www.ouw.co.uk, or contact Open University Worldwide, Michael Young Building, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom, for a brochure: tel. +44 (0)1908 858785, fax +44 (0)1908 858787, e-mail ouwenq@open.ac.uk

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 1997. Second edition 2004. Reprinted 2005.

Edited, designed and typeset by The Open University, using the Open University TEX System.

Printed in the United Kingdom by Thanet Press Ltd, Margate.

ISBN 0 7492 6662 7

2.2

Contents

Study guide

This chapter contains six sections, which are intended to be studied consecutively, and an appendix. Sections 3 and 4 should take two to three hours of study each. The other sections are shorter.

Sections 3, 4 and 5 contain only computer-based work, and will require the use of the computer and Computer Book D. Subsection 3.1 also requires the use of an audio cassette player.

The pattern of study for each session might be as follows.

Study session 1: Sections 1 and 2.

Study session 2: Section 3

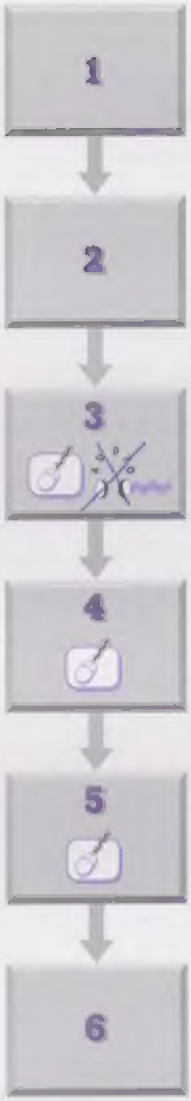
Study session 3: Section 4

Study session 4: Sections 5 and 6.

Before studying this chapter, you should be familiar with the following topics, ~~which are covered in the software package StatsAid:~~

- the standard deviation of a batch of data;
- scatterplots.

Both topics are covered in the course MU120.



Introduction

Many commonplace phenomena exhibit variability – the lengths of pregnancies, the heights and weights of adults, the amounts people earn, annual rainfall, the times between major earthquakes or other natural disasters, and so on. What they all have in common is that none can be predicted with certainty. For instance, when the next major earthquake will occur is unknown; sometimes two or three occur within a few months of each other, while at other times there will not be one for a year or more – the times between major earthquakes vary. And although human pregnancies are supposed to last 40 weeks, few babies arrive exactly 'on time' – the lengths of pregnancies vary. Babies' birthweights vary too. And some men and women are taller than others.

For each of these situations, we might collect some relevant data, calculate the mean, and use this to estimate, for instance, when the next major earthquake might occur, or the average height of all men, or the average weight of a baby at birth. But this will not tell us anything about how inaccurate our estimate might be, and it could be important to know this also.

There are many situations where knowledge of how physical measurements vary is important. Clothing designers need to know about the variability of many measurements such as, for example, chest, bust, waist and hip girth, arm length and inside and outside leg length; and they also need to know how these measurements are related to each other. Manufacturers need to know the distributions of such measurements in order to make informed decisions about the numbers of each size to produce.

Because humans vary in size, the distributions of physical measurements are also of great importance in the design of equipment for human use. The first systematic attempt to develop principles for equipment design and to collect data and apply the information obtained took place during the Second World War. This knowledge was used most effectively in the design of aircraft cockpits, oxygen masks and clothing.

In the late 1940s, investigations were carried out in the United States into the use of all types of equipment, both civilian and military, with a view to improving safety. Among other things, data were collected on measurements such as sitting eye height, arm reach and leg length for bus and truck drivers. Poor design increases the risk of accidents: for example, arms tire if the steering wheel is too far from the body, and if the foot pedals are too close, then excessive bending of the ankles is required. Effective use of equipment depends on good design, and this requires knowledge of the physical dimensions of those likely to operate the equipment.

Ideally, all men and women should be able to operate all equipment: car manufacturers, for instance, want their vehicles to be usable by everyone. And, in practice, provided that the overall dimensions are chosen correctly, adjustable devices can accommodate almost everyone. However, relevant data on the distribution of the physical dimensions of people are essential for the designers if the goal of universal operability is to be achieved.

The design of equipment such as aircraft cockpits or driving cabs or cars is very complicated, so, in order to introduce and develop ideas about how to model variation, we shall consider the much simpler problem of how to model the heights of men.

This chapter is about modelling the variation observed in data; and it is concerned primarily with one particular model for variation called the *normal distribution*. In Section 1, we begin by discussing the general problem of how to model the variation observed in a sample of data, and then the normal distribution is introduced. In Chapter D1, the geometric distribution was introduced as a model for the variation observed in the number of trials of an experiment needed to obtain a success. You saw that, in general, choosing a geometric distribution to model a particular situation involved fixing the value of one parameter: p, the probability of success in each trial of the experiment. Choosing a normal distribution to model the variation observed in data involves fixing the values of two parameters. In Section 2, we discuss what these parameters represent and how data are used to select suitable values for them. The ideas introduced in the first two sections are used in Sections 3 and 4 to select models for the variation in a number of phenomena – in each case, the model is chosen to reflect the variation observed in a sample of data.

There is considerable evidence that, within a country, human physical measurements change over a period of time. For example, records show that the average height of American soldiers was 0.7 inches greater in the Second World War than in the First World War. And in the 1960s, the trend for average heights to increase was noted for many other Western countries and also for Japan. In Section 4, you are invited to explore a large data set collected at the beginning of the 20th century on heights of fathers and sons, and to investigate whether, in general, the sons were taller than their fathers. Finally, the properties of normal distributions are investigated in Section 5 and summarised in Section 6.

In Section 3, you will be introduced to the data analysis software *OUStats for MST121*, which from now on we shall call *OUStats*. Mathcad and OUStats are both *Windows*-based software packages, and some operations are carried out in a similar way for the two packages. So you can use your experience with Mathcad to help you when learning to use OUStats.

1 Models for variation

In Chapter D1, you saw how to model the variability in one particular quantity – the number of rolls of a die needed to obtain a six. By assuming that the die is 'fair' – that is, that each score is equally likely to occur on each roll of the die – we developed a model for the variation in the number of rolls required to obtain a six: this model is the geometric distribution. The model specifies the probability of obtaining a six on each possible number of rolls: $1, 2, 3, \ldots$. It gives a good indication of the likely variation in the number of rolls that may be needed to obtain a six on different occasions. This distribution is illustrated in Figure 1.1(a).

In Section 2 of Chapter D1, you saw that simulations, based on the assumption that the six possible scores on each roll are equally likely to occur, gave rise to frequency diagrams which were similar in shape to the probability distribution in Figure 1.1(a). An example is shown in Figure 1.1(b), showing that the model seems to be a good one for what actually happens.

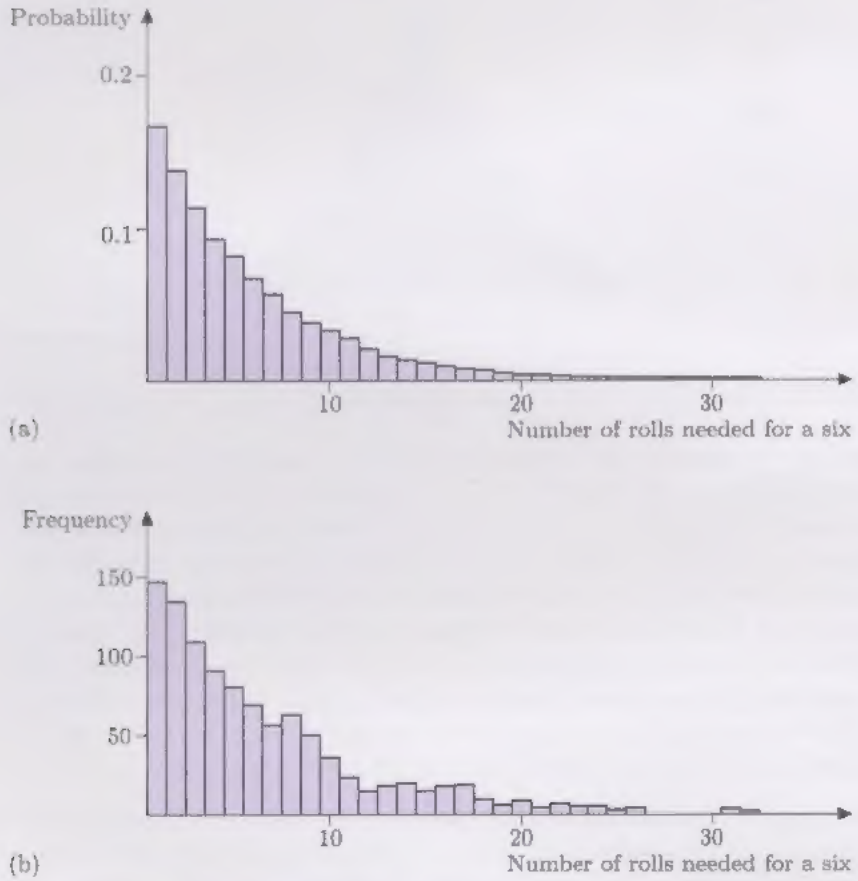


Figure 1.1 (a) A geometric distribution with $p = \frac{1}{6}$ (b) The results of a simulation

Now consider how we might model the variation in men's heights. Since it is not practicable to measure the heights of all men, the first step is to obtain some data. It proved surprisingly difficult to find data on the heights of individual men. Data on the distributions of human measurements can be of considerable commercial value, so many organisations that collect such data are reluctant to allow their

publication. However, some searching turned up an old data set which contains the heights of 1000 Cambridge men in 1902. These heights are a sample of the heights of all Cambridge men in 1902, and will be used to illustrate how the variation observed in men's heights may be modelled.

In 1902, inches were used in the UK to measure heights, rather than metric units.

Figure 1.2 shows a frequency diagram for the heights of 1000 Cambridge men in 1902. The heights were recorded to the nearest inch; so, for instance, all men strictly between 69.5 and 70.5 inches tall were recorded as 70 inches tall and are represented by the bar centred on 70 inches. (Rounding of heights was used: a height measured as exactly 69.5 inches was recorded as 70 inches; a height measured as exactly 70.5 inches was recorded as 71 inches.) These data are from an article in which various measurements taken on convicted criminals were compared with corresponding measurements for men in the general population. Around this period, there was considerable interest in discovering whether there were differences between various physical measurements for criminals and for the rest of the population. In the article, the Cambridge men were taken to be representative of the general population.

Source: W. R. MacDonell, 'On criminal anthropometry and the identification of criminals', *Biometrika* 1 (1902) 177–227.

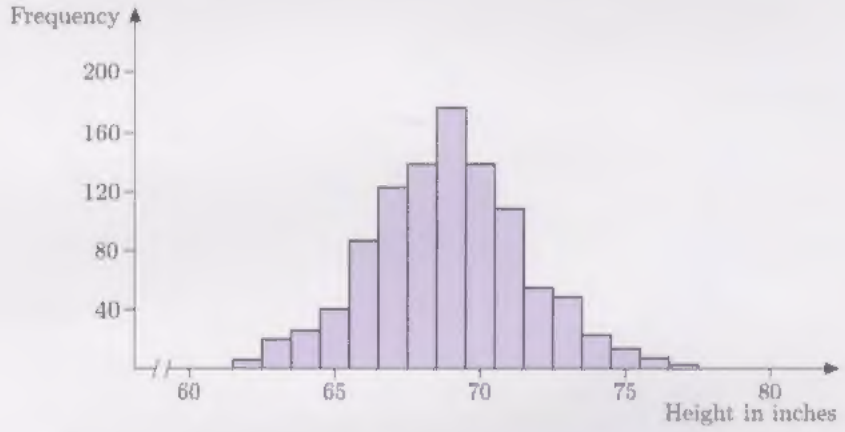


Figure 1.2 The heights of 1000 Cambridge men

Clearly, the geometric distribution in Figure 1.1(a) is not a suitable model for the variation in the heights of Cambridge men in 1902, a sample of whose heights is represented in Figure 1.2. First, the shapes of the two diagrams are different. We would not expect the frequency diagram for a sample drawn from a population with a distribution like that in Figure 1.1(a) to look like that in Figure 1.2. And secondly, the geometric distribution is a discrete distribution – it is a model for a variable which can take only the separate, distinct values $1, 2, 3, \ldots$; whereas a height may take any value in an interval of values – in this case, between about 60 inches and 80 inches – that is, it is a *continuous* variable.

In Chapter D1, the geometric distribution was obtained as a model for the variation in the number of rolls of a die needed to obtain a six by assuming that each score on a die is equally likely to occur on each roll. There is no comparable assumption about heights of men that would allow us to use probability theory to develop a model for the variation in the heights of men. So a different approach is needed in this case.

Instead of using theory to develop a model (as we did when considering the number of rolls of a die needed to obtain a six), we shall choose a model to 'fit' the data; this model will be one of a collection of standard models. If the model that we choose produces a 'good fit', then simulations using the model should give rise to frequency diagrams which are similar in

shape to the frequency diagram for the original data. In particular, the two diagrams should have peaks in similar positions and have a similar spread of values. So an essential feature of a model for the heights in Figure 1.2 is that its probability diagram should be broadly similar in shape to the frequency diagram. In addition, the mean value predicted by the model should be roughly equal to the mean of the data, and the range of values which, according to the model, are likely to occur should correspond roughly to the range of values in the data. Finally, the model must allow all values in an interval to occur, not just a discrete set of values – that is, a *continuous* model is required.

Activity 1.1 Properties of the data

Describe the shape of the frequency diagram in Figure 1.2. Use the diagram to estimate very roughly the mean height of the men. Over what range are the heights of the men spread?

Comment

The frequency diagram is roughly symmetrical about a single peak or **mode**; it is **unimodal**. The frequencies are low on the left side of the diagram, rise steadily to reach a maximum for heights of 69 inches, then decrease towards the right. The diagram could be described as approximately 'bell-shaped'. Since the diagram is roughly symmetrical, the mean height is approximately 69 inches, the 'centre' height. The heights of the men range from approximately 61.5 inches to approximately 77.5 inches.

So what would the probability diagram of a model for the heights of Cambridge men in 1902 look like? Ideally, it should possess all the properties of the frequency diagram just described. This suggests that it should be 'bell-shaped' with a peak at 69 inches and it should allow heights between, say, 60 and 78 inches. One possibility is shown in Figure 1.3.

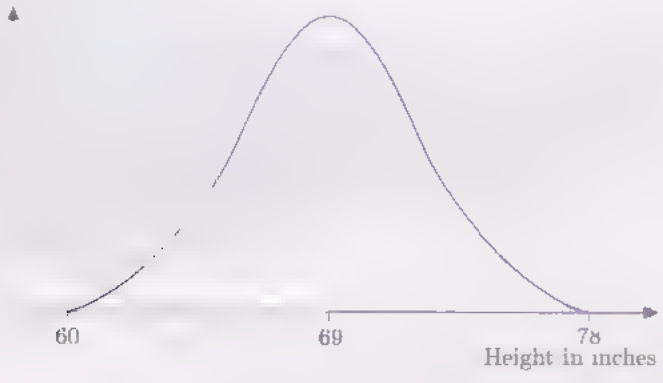


Figure 1.3 A possible model

Notice that in this figure the model has been represented by a curve, rather than by a series of bars (as in the discrete probability diagram for a geometric distribution). It is usual to use a curve when modelling a continuous variable. But how is such a model to be interpreted? And how can it be used to estimate the proportion of (say, all Cambridge men who were over six feet tall, for example, or the proportion who were between 69 and 71 inches tall?

Essentially, areas under the curve are used to represent proportions or probabilities. You will see how this is done later in this section when we return to the question of the interpretation of the model. The scale on the vertical axis, which is omitted in Figure 1.3, is chosen so that areas represent proportions. In practice, as you will see later, you do not need to worry about including the vertical scale when using a curve as a model: often, we shall not even include the vertical axis. The essential point at this stage is that the curve should be similar in shape to the frequency diagram for the data.

Activity 1.2 Models for variation

Figure 1.4 contains frequency diagrams for four data sets.

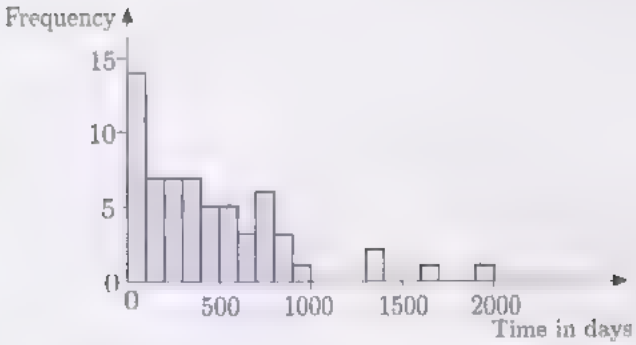
Figure 1.4(a) represents the times between 'major' earthquakes from December 1902 until March 1977. Any earthquake whose magnitude was at least 7.5 on the Richter scale or in which over a thousand people were killed has been included. There were 63 major earthquakes according to these criteria, so there are 62 times between earthquakes in the data set.

Figure 1.4(b) represents the wing lengths of 252 first-year meadow pipits netted in southern Scotland in the autumn of 1991. We shall investigate these data further in Chapter D4.

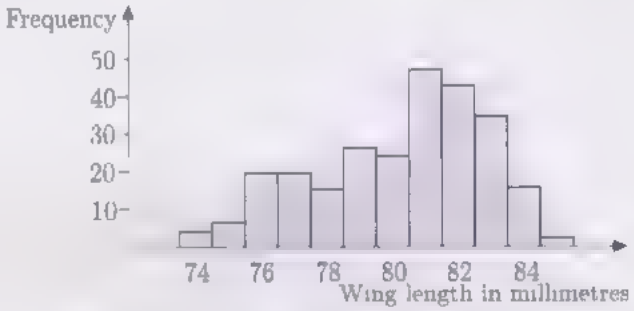
Figure 1.4(c) represents the durations of 106 eruptions of the 'Old Faithful' geyser in Yellowstone National Park, USA in August 1978.

Figure 1.4(d) represents the gross weekly earnings (including overtime payments) in April 1994 of 150 women working full-time.

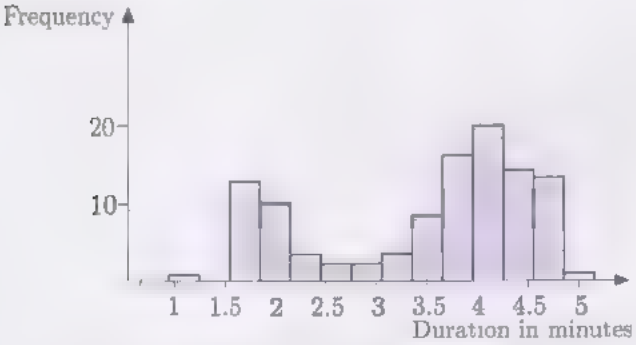
Sources (a) *S237 The Earth: structure, composition and evolution*, Block 2; (b) T. W. Dougall, Post-juvenile moult in the Meadow Pipit *Bird Study*, 41 (1994), 137–142



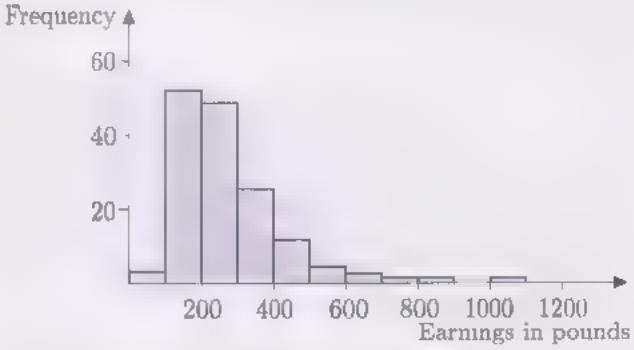
(a) Times between major earthquakes



(b) Wing lengths of meadow pipits



(c) Durations of eruptions of Old Faithful geyser



(d) Weekly earnings of women employees

Figure 1.4 Frequency diagrams for four data sets

For each of the four data sets, briefly describe the shape of its frequency diagram and sketch a curve which you think might reasonably model the variability in the data. Mark on your sketch any key values (as was done in Figure 1.3). (Do not worry about the scale on the vertical axis.)

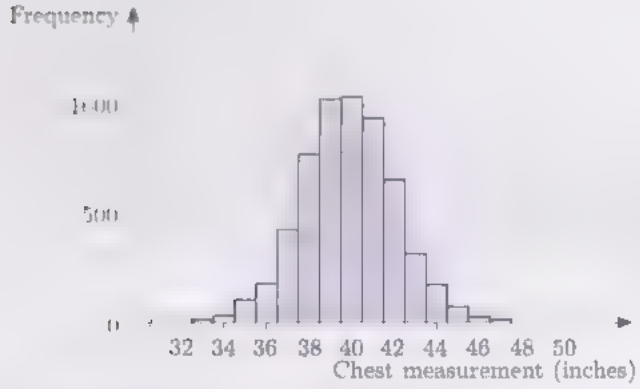
A solution is given on page 41.

Comment

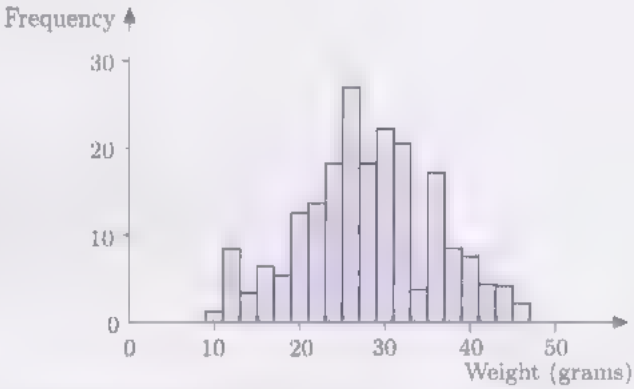
You will probably have noticed that three of these four frequency diagrams have one tail of values much more extended than the other. (In fact, the frequency diagram for the earthquakes data has no left tail at all.) A data set or diagram which is asymmetric because of a long tail is said to be **skewed**. If it has a long right tail, then it is said to be **right-skew**; if it has a long left tail, it is said to be **left-skew**. The terms 'right-skew' and 'left-skew' are useful ones for describing the shape of a diagram or data set. Figure 1.4(a) and Figure 1.4(d) are right-skew, and Figure 1.4(b) is left-skew.

Activity 1.3 Similar models

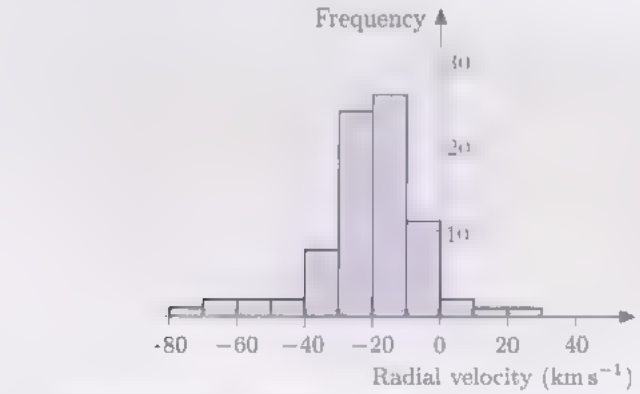
Figure 1.5 contains frequency diagrams for four data sets.



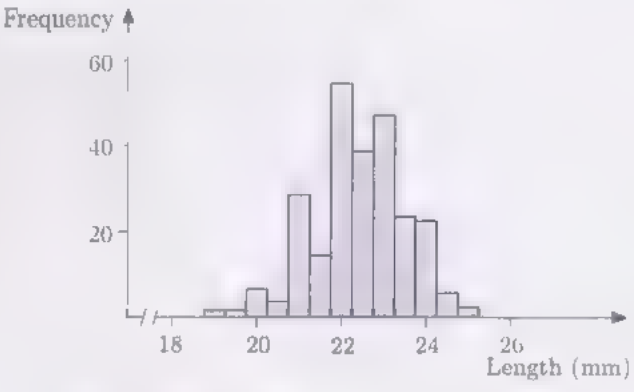
(a) Scottish soldiers' chest measurements



(b) Weights of Irish dipper nestlings at 6–8 days



(c) Radial velocities of stars



(d) Lengths of cuckoo eggs

Figure 1.5 Frequency diagrams for four more data sets

Sources: (a) S. M. Stigler, *The History of Uncertainty before 1900* (Belknap Press, Cambridge, MA, 1986); (b) John O'Halloran, Patrick Smiddy and Barry O'Mahony, 'Biometrics, growth and sex ratios amongst Irish Dippers', *Ringing and Migration* 13 (1992) 152–161; (c) R. J. Trumpler and H. F. Weaver, *Statistical Astronomy* (University of California Press, Berkeley, CA, 1953); (d) O. H. Latter, 'The egg of *Cuculus canorus*', *Biometrika* 1 (1902) 164–176.

Figure 1.5(a) represents the chest measurements of 5732 Scottish soldiers.

Figure 1.5(b) represents the weights of 198 Irish dipper nestlings aged 6–8 days.

Figure 1.5(c) represents the radial velocities of 80 bright stars in a certain region of the sky. (The *radial velocity* of a star is the velocity with which the star appears to be approaching the Earth or, more precisely, the component of its velocity towards the Earth. If the radial velocity of a star is negative, then the star is moving away from the Earth.)

Figure 1.5(d) represents the lengths of 243 cuckoo eggs.

(i) These frequency diagrams are all similar in shape. How would you describe their common shape?

(ii) For each data set, sketch a curve which you think might reasonably model the variability in the data. Mark any key values on your sketches. (Do not worry about the vertical scale.)

Comment

(i) Each of the diagrams is unimodal and roughly symmetrical about its peak. In each diagram, the frequencies are low on the left-hand side, rise steadily to reach a maximum in the centre, and then decrease towards the right. Apart from some jaggedness in Figure 1.5(b) and Figure 1.5(d), the diagrams are similar in shape to the frequency diagram of heights of Cambridge men in Figure 1.2; they are all roughly 'bell-shaped'. For each data set, the peak is in a different position and the spread of the values is different, but all the diagrams have the same underlying shape.

(ii) Sketches of possible models are shown in Figure 1.6. The jaggedness in some of the frequency diagrams could simply be due to chance variation. In Section 3, you will be asked to investigate whether this might be the case, using the statistics software.

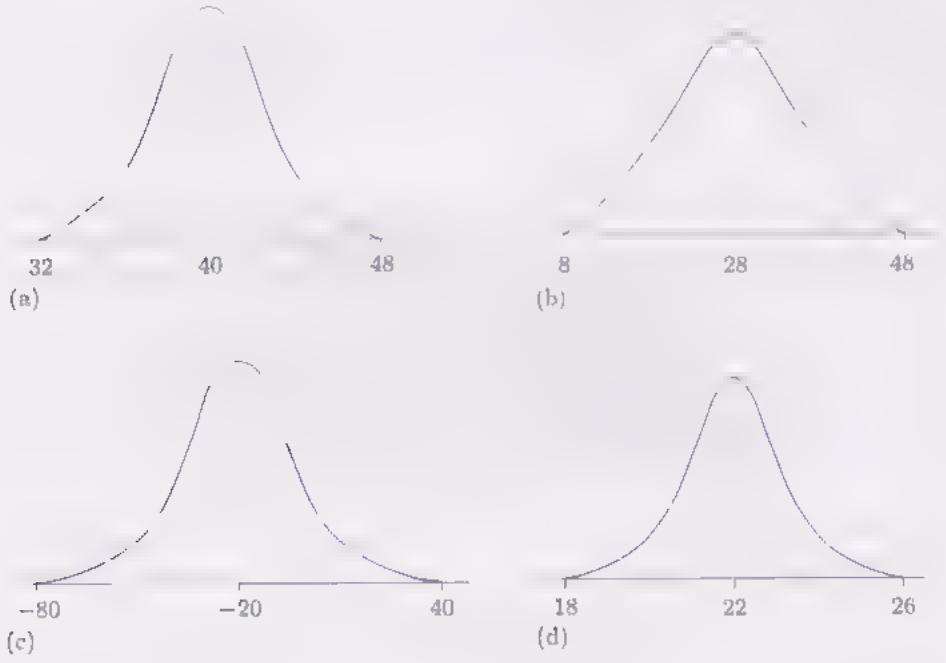


Figure 1.6 Possible models

The normal distribution

The frequency diagrams in Figures 1.2 and 1.5 are all essentially the same shape – in each case, a 'bell-shaped' curve would seem to be a good model for the variation in the data. A model which has this shape is the *normal distribution*. In fact, there is a whole family of normal distributions, *all with the same basic shape*. It is possible to choose a member of this family so that its peak is in whatever position is required and with whatever spread of values is needed. The equation of a typical normal curve is $y = f(x)$, where

As used here, the word 'normal' is a technical term and does not have its everyday meaning.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (-\infty < x < \infty). \tag{1.1}$$

You do not need to remember this formula

Notice that (1.1) contains two parameters, μ and σ. Each pair of values for μ and σ gives a different member of the family of normal distributions. In some texts, symbols other than μ and σ are used for the parameters of a normal distribution (a and b, for example). However, μ and σ are the symbols most commonly used, and we shall use them in this course.

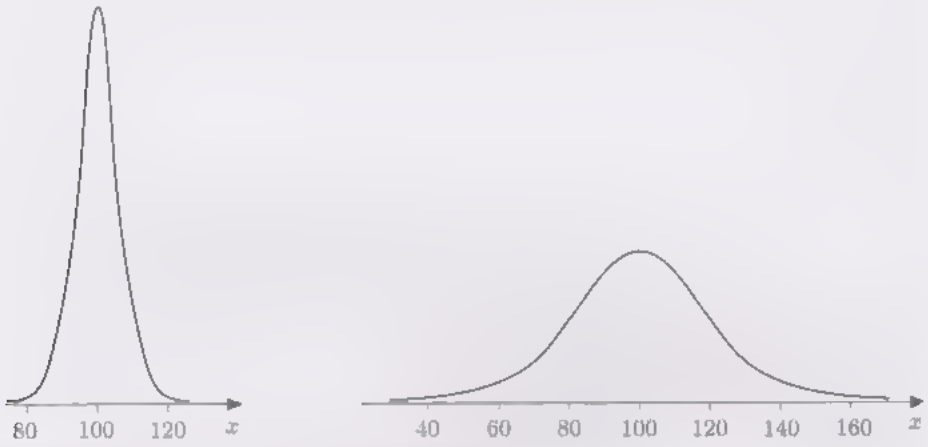
μ and σ are the Greek lower-case letters mu and sigma.

The function f is known as the **probability density function** of the distribution. Note that any value at all for x produces a corresponding value for $f(x)$, so the normal distribution is a continuous model. A sketch of this curve is shown in Figure 1.7.

Figure 1.7 A typical normal curve

The normal curve is symmetrical about the line $x = \mu$, so the parameter μ gives the location of the 'centre' of the curve. Since the heights of Cambridge men in Figure 1.2 are distributed roughly symmetrically about 69 inches, we might use a normal distribution with parameter $\mu = 69$ to model the variation in the heights of Cambridge men. Similarly, a normal distribution with $\mu = 40$ might be used to model the variation in Scottish soldiers' chest measurements shown in Figure 1.5(a).

The parameter σ governs the spread of the values that, according to the model, are most likely to occur: in general, for fixed horizontal and vertical scales, a small value of σ produces a tall narrow 'bell', and a large value of σ produces a short wide 'bell'. Two examples are shown in Figure 1.8. The roles of the two parameters and the reason for the use of the symbols μ and σ are discussed further in Section 2.



(a) Smaller σ (b) Larger σ

Figure 1.8 Two normal curves (same vertical and horizontal scales)

In Section 2, we shall discuss how to choose values for μ and σ in order to select an appropriate normal distribution as a model for the variation observed in a data set; and in Section 3, you will have the opportunity to fit normal models to a number of sets of data using the statistics software.

A normal curve is symmetrical about its peak and falls away on either side, gradually approaching the x-axis. Note, however, that the curve never actually meets the x-axis, unlike the sketches of possible models drawn in Activity 1.3. In this model, any number between $-\infty$ and $+\infty$ is possible, although values far from the peak are very unlikely. You might wonder whether it is reasonable to model measurements such as height, length and chest size, which cannot be negative, using a distribution which allows negative values. The point to remember is that a normal distribution is *only a model*: in practice, when fitting a normal model to such data, the proportion of values that, according to the model, are negative is so small that it may be ignored. You can see this for yourself in the computer session in Section 3.

Using a curve to model variability

To see how a normal curve can be used to model variability, we return to the data on the heights of 1000 Cambridge men which were represented in the frequency diagram in Figure 1.2. This diagram is repeated below as Figure 1.9(a).

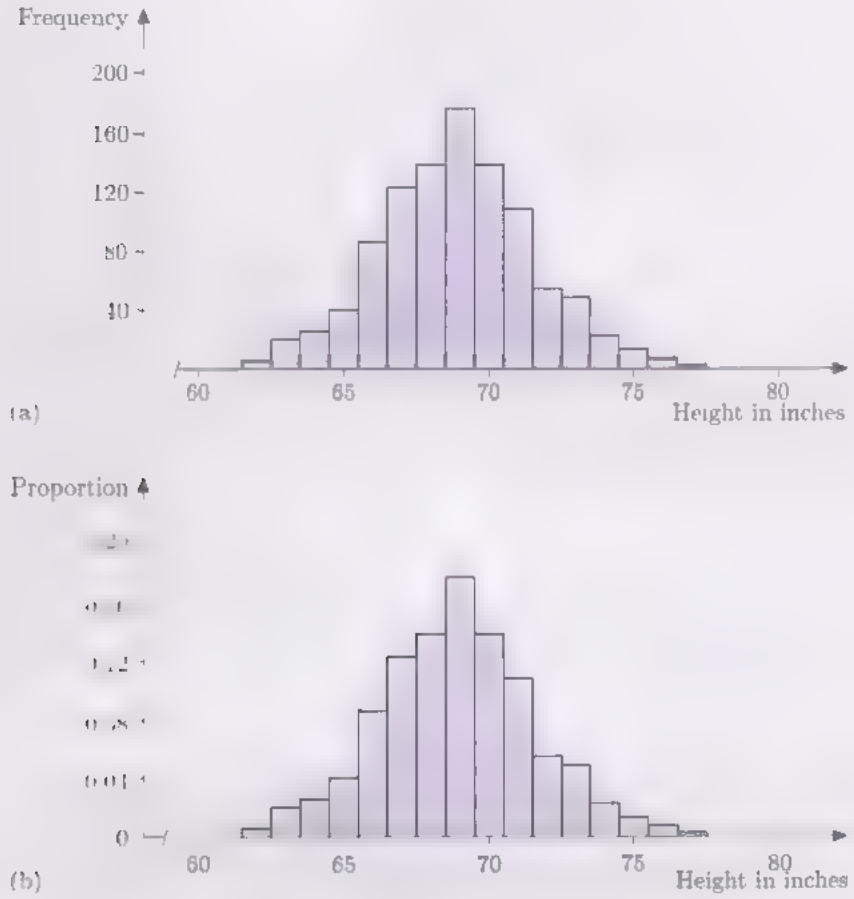


Figure 1.9 (a) A frequency diagram (b) A histogram

Figure 1.9(b) shows a *histogram* for the same data. While in a frequency diagram the *heights* of the bars represent the various frequencies, in a histogram it is the *areas* of the bars which are proportional to the frequencies. To obtain the histogram in Figure 1.9(b) from the frequency diagram in Figure 1.9(a), the scale on the vertical axis was adjusted so that the sum of the areas of all the bars is equal to 1. In all other respects, the two diagrams are identical. Since the total area of the bars in the histogram is equal to 1, this means that, for example, the area of the bar centred on 70 inches, which represents heights between 69.5 and 70.5 inches, is equal to the proportion of the 1000 men who were between 69.5 and 70.5 inches tall. In Figure 1.9(a), the height of this bar is 139, since 139 of the 1000 men had heights in this range; in Figure 1.9(b), the area of the corresponding bar is 0.139, since this is the proportion of men whose heights were in this range.

Suppose now that an appropriately chosen normal curve is superimposed on the histogram in Figure 1.9(b); this is shown in Figure 1.10 overleaf. (You will learn how to choose an appropriate normal curve in the next section.)

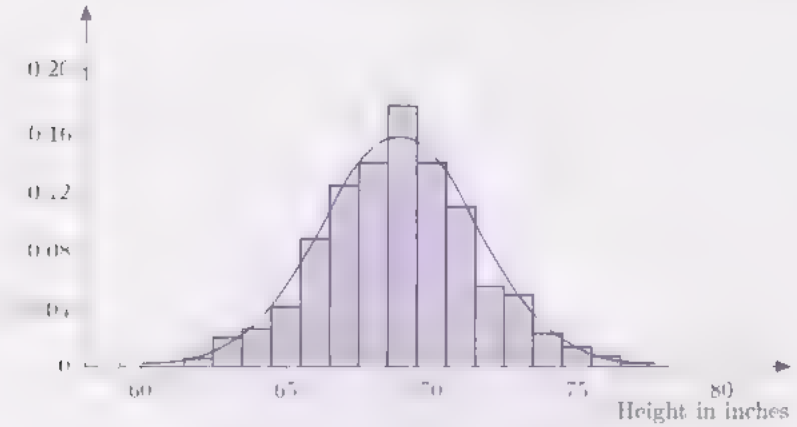


Figure 1.10 The histogram and a fitted normal curve

The total area between the normal curve and the x-axis may be obtained as follows.

The area between the normal curve and the x-axis from $x = -N$ to $x = N$ is given by

$$\int_{-N}^{N} f(x)\,dx,$$

where $f(x)$ is given by formula (1.1). This integral can be calculated for any $N > 0$, and its value approaches a limiting value as N becomes larger and larger. The total area between the normal curve and the x-axis is given by this limiting value. We write

$$\text{area} = \int_{-\infty}^{\infty} f(x)\,dx = \lim_{N \to \infty} \int_{-N}^{N} f(x)\,dx.$$

Some information about several results involving integrals with infinite limits that are stated without proof in this chapter are included in an appendix (page 40).

For *any* normal curve this area is always equal to 1, no matter what are the values of the parameters μ and σ. So the total area under any normal curve is equal to the total area of the bars in a histogram, provided the histogram is drawn so that the total area of the bars is equal to 1.

As you can see, the area under the part of the curve in Figure 1.10 from 69.5 to 70.5 is approximately equal to the area of the corresponding bar in the histogram. These areas are shown in Figure 1.11.

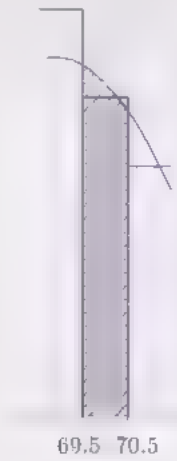


Figure 1.11 The area under the curve and the area of the bar

The shaded area under the curve in Figure 1.11 is, in fact, equal to 0.141 to three decimal places. (In Section 3, you will learn how to use the statistics software to find this area.) So the area under the curve between 69.5 and 70.5 is approximately equal to the proportion of men in the sample who were between 69.5 and 70.5 inches tall — 0.141 in the model compared with 0.139 in the sample.

The histogram represents the heights of just 1000 Cambridge men in 1902. If we assume that these were a sample from the population of all men in Cambridge, then the normal curve provides a model for the variation in the heights of *all* Cambridge men in 1902. The area under the curve between the two heights 69.5 inches and 70.5 inches represents the proportion of *all* Cambridge men that were, *according to the model*, between 69.5 and 70.5 inches tall.

We can generalise the interpretation of areas under the normal curve as follows. An area under the curve between two heights, a inches and b inches say, represents the proportion of *all* Cambridge men that were, *according to the model*, between a inches and b inches tall. This is illustrated in Figure 1.12.

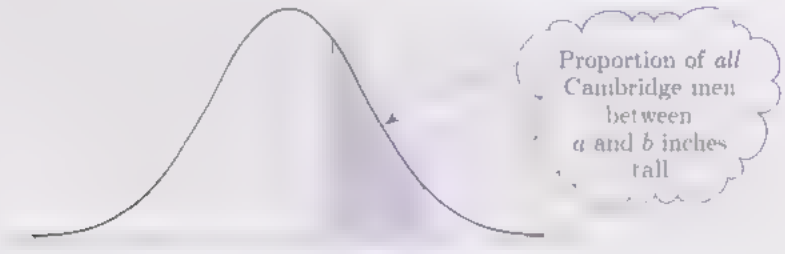


Figure 1.12 The model

The shaded area could be found by evaluating the integral

$$\int_a^b f(x)\,dx,$$

where $f(x)$ is given by formula (1.1).

Since the area represents the proportion of Cambridge men that were between a and b inches tall, it follows that if a man had been selected at random from all the men in Cambridge in 1902, then the probability that his height would have been somewhere between a inches and b inches is also given by this area. This gives us a second interpretation of the area. So, for instance, if a man had been selected at random from all Cambridge men, then the probability that his height would have been between 69.5 and 70.5 inches is given by the area under the curve between 69.5 and 70.5.

Activity 1.4 Interpreting the model

Figure 1.13 contains three sketches of the particular normal curve used to model the heights of all Cambridge men in 1902. For each sketch, describe in words what the shaded area represents.

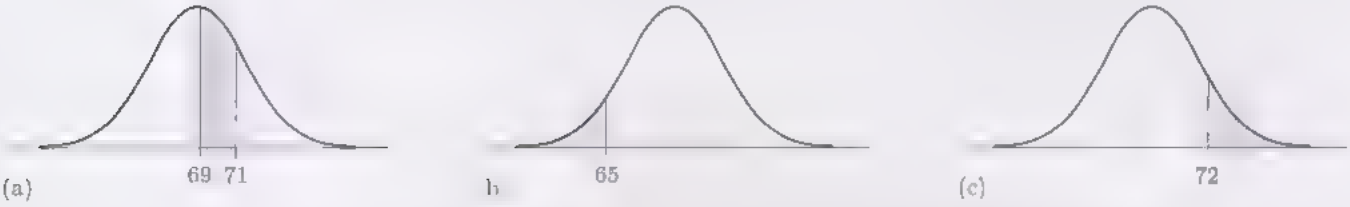


Figure 1.13 Three areas under a normal curve used to model the heights of Cambridge men in 1902

A solution is given on page 41.

You might wonder why we should use a model to estimate the proportion of Cambridge men between, say, 69 and 71 inches tall. Surely it would have been better to measure all the men and find the actual proportion whose height was in this range? In theory, it would have been better, but in practice, this would not have been a realistic course of action to take. The work involved in measuring the heights of 1000 men must have been very considerable. To measure the height of every adult man in Cambridge would have been a very difficult, if not impossible, task.

This sort of situation is a common one. Information about an entire population is needed (perhaps for developing marketing or manufacturing plans, or for making decisions about the future provision of health care or education, or perhaps simply to gain knowledge), but it is impractical or too costly to examine (or interview or measure, as appropriate) every member of the population. A practical approach is to examine a sample drawn from the population.

If the sample is suitably selected (ideally, it should be a random sample that is, a sample drawn in such a way that every member of the population has an equal chance of being chosen), then we can use statistics to infer information about the population from what we find out about the sample. And in order to make inferences about the population, we frequently need to develop a model for the variability in the population as a whole.

What we have been doing is seeking a model for the variation in the heights of all Cambridge men which fits the data available for the sample of 1000 men. This model can then be used to estimate the proportion of all Cambridge men whose height is in any given range. Such proportions may be found by calculating appropriate areas under the curve used to model the variation in heights.

A little history

The normal distribution is one of the most frequently used models for variation; its specification given in (1.1) was obtained first by Abraham de Moivre (1667–1754). He derived it as an approximation while endeavouring to compute probabilities in various games of chance. After the existence of a belt of asteroids between Mars and Jupiter was discovered at the beginning of the 19th century, the study of the motions of the asteroids was the subject of considerable scientific activity. One of the very early applications of the normal distribution was in describing the variability in errors of measurement of the motions of such asteroids.

The normal distribution is also known as the *Gaussian distribution*, after the German mathematician and astronomer Karl Gauss (1777–1855), who played a leading role in demonstrating the usefulness of the distribution as a model for measurement errors in general in the physical sciences. Since then, the distribution has been used successfully to model variability in many phenomena in other fields of study. Indeed, for a time in the 19th century, there was a school of thought which maintained that the normal distribution was the only model necessary to describe variability, and a great deal of effort was put into trying to prove that this was the case.

Although the normal distribution is a very useful model, it is not universally applicable. You have already seen this: in Activity 1.2, you met four examples where a normal curve would not have been an appropriate model. But as you will see in Chapter D3, in certain circumstances the normal distribution may still be used in a statistical investigation, even when it is not a suitable model for the variability in the available data, and this is the reason for its importance in statistics.

Summary of Section 1

In this section, you have seen how a curve may be used to model the variation in various quantities, such as men's heights, women's earnings, the times between major earthquakes and the wing lengths of meadow pipits. The frequency diagrams for several of the data sets had the same 'bell-shape'. A distribution was introduced as a model for variation of this sort – the normal distribution. You saw how areas under a normal curve can be used to answer questions about the population being modelled.

2 Choosing a normal model

In Section 1, you saw that the equation of a normal curve contains two parameters, denoted μ and σ. In order to choose a specific normal curve to model the variability in some quantity - such as, for example, the heights of Cambridge men or the weights of Irish dipper nestlings - you need to specify values of these parameters. The purpose of this short section is to discuss what they represent and how to select values for them.

In Section 1, it was observed that the parameter μ provides the location of the 'centre' of the curve, and the parameter σ governs the spread of values that are most likely to occur. In fact, the parameter μ is the mean of the distribution and σ is a measure of spread called the standard deviation. The symbols μ and σ are used generally to denote the mean and standard deviation of a distribution. In practice, probability distributions are used to model the variation in populations, and so the symbols μ and σ are also used to denote the mean and standard deviation of a population.

μ (mu) is the Greek letter 'm' and σ (sigma) is the Greek letter 's'. There is therefore a first-letter link (albeit in a different alphabet) between the symbol for the parameter and what it stands for.

The use of the symbols μ and σ for the mean and standard deviation of both distributions and populations does not usually lead to any confusion, because of the close connection between a population and a probability distribution used to model the variation in the population. In fact, the mean of a probability distribution is sometimes called the *population mean* and the standard deviation of a probability distribution is sometimes called the *population standard deviation*. The population mean μ and the population standard deviation σ are called *population parameters*.

One of the data sets in Section 1 consisted of the heights of 1000 Cambridge men in 1902. Suppose that we wish to use a normal curve to model the variability in the heights of all Cambridge men in 1902. If the model is to reflect accurately the distribution of heights in the population, then the parameter μ should be equal to the mean height of all Cambridge men - the population mean - and the parameter σ should be equal to the standard deviation of the heights - the population standard deviation. So we ought to use the population mean and the population standard deviation as values for the parameters of the normal model.

But we do not know the mean and standard deviation of the population. However, we can use the sample of 1000 heights drawn from the population to calculate the corresponding *sample statistics* - the *sample mean*, which is denoted $\bar{x}$, and the *sample standard deviation*, denoted s. The sample mean $\bar{x}$ may then be used as an estimate for the population mean, and the sample standard deviation s as an estimate for the population standard deviation. The approach that we shall adopt in Section 3, when we use the statistics software to fit a normal curve to data, is to use the values of the sample mean $\bar{x}$ and the sample standard deviation s to estimate the population mean μ and the population standard deviation σ, respectively. So the values of the parameters of the normal model that we fit will be given by $\bar{x}$ and s.

In general, letters from the Roman alphabet are commonly used for *sample* statistics, and letters from the Greek alphabet are used for *population* parameters.

In this section, we shall discuss briefly how the population parameters μ and σ are defined, and how the sample statistics $\bar{x}$ and s are calculated. The population mean μ and the sample mean $\bar{x}$ are discussed in Subsection 2.1; the population standard deviation σ and the sample standard deviation s are discussed in Subsection 2.2. This section contains quite a lot of detail that you are not expected to remember. For instance,

in this course, you will not be expected to calculate the mean and standard deviation of a probability distribution. When these are required (in Chapter D3, for example), you will be given their values. But after working through this section, you should make sure that you understand the distinction between population parameters such as μ and σ and sample statistics such as $\bar{x}$ and s, and that you can calculate the mean and standard deviation of a small data set using a calculator. In Section 3, you will learn how to use the statistics software to calculate the sample mean and sample standard deviation.

2.1 The mean

Many of the ideas in this subsection will be familiar to you. For instance, you should already be familiar with the calculations involved in finding the mean of a sample of data (perhaps from studying MU120 or the *Revision Pack*). And the mean of a discrete distribution (such as the geometric distribution) was introduced in Section 4 of Chapter D1. Before discussing how the mean of a continuous distribution such as the normal distribution is defined, we shall review briefly how the mean of a sample of data is calculated.

The sample mean

Suppose that a sample of n observations is taken from a population. (For example, these might be the heights of n men.) We shall denote these observations by $x_1, x_2, \ldots, x_n$. The sample mean $\bar{x}$, which is defined to be the sum of the observations divided by the number of observations, is given by the formula

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Sometimes, many of the recorded values occur several times. In this case, the data are often placed in a table showing the number of times that each different value was observed — its **frequency**. For example, Table 2.1 contains the data on the heights of 1000 Cambridge men which were represented in Figure 1.2.

Table 2.1 Heights of 1000 Cambridge men

Height in inches	Frequency
62	3
63	21
64	24
65	1
66	57
67	122
68	35
69	179
70	183
71	107
72	77
73	17
74	22
75	2
76	5
77	1
	1000

You might like to think about why the two formulas are equivalent.

To find the mean height when data are given in this form, we use the equivalent formula

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i = \frac{1}{n}(x_1 f_1 + x_2 f_2 + \cdots + x_k f_k),$$

where $x_1, x_2, \ldots, x_k$ are the different values observed in the sample and $f_1, f_2, \ldots, f_k$ are their respective frequencies; the sample size n is equal to $f_1 + f_2 + \cdots + f_k$, the sum of the frequencies.

For the sample of 1000 heights of Cambridge men,

$$\bar{x} = \frac{1}{1000}(62 \times 3 + 63 \times 20 + \cdots + 77 \times 1) = 68.872.$$

So the sample mean is approximately 68.9 inches. The term in brackets is the sum of the heights of all 1000 men: 3 men were 62 inches tall, so 62 is included 3 times (62×3); 20 men were 63 inches tall, so 63 is included 20 times (63×20); and so on.

In fact, the heights given in Table 2.1 were recorded rounded to the nearest inch, so that a man recorded as 62 inches tall may have been as much as half an inch shorter or taller than 62 inches. All men who were at least 61.5 inches tall but less than 62.5 inches tall were grouped together and recorded as being 62 inches tall. Similarly, all men between 68.5 and 69.5 inches tall were grouped together and recorded as 69 inches tall. So it is possible, even likely, that the value of the sample mean calculated above is not exactly equal to the actual mean height of the 1000 men. However, given the way the data have been recorded, it is the best that we can do: the midpoint of each interval of heights has been used instead of the exact heights of the men in each group.

Data are often summarised and reported by grouping the values and recording the frequencies for each group. In this case, the midpoint of each interval is used in calculating statistics such as the sample mean. This was done implicitly when calculating the mean height of the Cambridge men.

Activity 2.1 Irish dipper nestlings

Table 2.2 Weights of Irish dipper nestlings

Weight in grams	Frequency
9–11	1
11–13	8
13–15	3
15–17	6
17–19	5
19–21	12
21–23	13
23–25	18
25–27	27
27–29	18
29–31	22
31–33	20
33–35	3
35–37	17
37–39	8
39–41	7
41–43	4
43–45	4
45–47	2
	198

The data which were used to draw the frequency diagram in Figure 1.5(b) are given in Table 2.2. These are the weights of 198 Irish dipper nestlings aged 6–8 days. What values would you use to calculate the mean weight of these nestlings? (Do not bother to go on to calculate the mean: leave this sort of lengthy calculation for the computer to do. You will learn how to do this in Section 3.)

Comment

The midpoints of the intervals are $10, 12, 14, \ldots, 46$, so these are the values that are to be used to calculate the sample mean. The sample mean is given by

$$\bar{x} = \frac{1}{198}(10 \times 1 + 12 \times 8 + \cdots + 46 \times 2).$$

The population mean

In Section 4 of Chapter D1, you were introduced to the idea of the mean of a probability distribution or the mean of a random variable X. In the situation discussed there, X was the number of rolls of a die needed to obtain a six, and you saw that this random variable has a geometric distribution. You also saw that, for a discrete distribution such as the geometric distribution, the mean of the distribution, which is denoted by μ, is given by the formula

$$\mu = \sum j \times P(X = j), \tag{2.1}$$

See Chapter D1, formula (4.4).

where the summation is over all values j that X can take. The probability distribution provides a model for the variability in the number of rolls of the die needed to obtain a six; so the mean of the distribution is the mean number of rolls predicted by this model.

Since a probability distribution may be used to model the variability in a population, the mean of the distribution is also sometimes called the **population mean**. (The population in the example above consists of the possible numbers of rolls of a die needed to obtain a six.)

For a discrete model, the population mean is given by formula (2.1). But clearly this is not appropriate for a continuous model, such as the normal distribution, in which all real values in an interval are possible. For a continuous model, variability is described by a curve, instead of a discrete set of probabilities, and areas under the curve represent probabilities.

The formula for the mean of a continuous distribution, which corresponds to formula (2.1) for the mean of a discrete distribution, is

$$\text{mean} = \int_{-\infty}^{\infty} x f(x)\, dx, \tag{2.2}$$

where f is the probability density function of the distribution (that is, $y = f(x)$ is the equation of the curve used as a model). Formula (2.2) is the continuous analogue of formula (2.1): an integral replaces the sum, and $f(x)$ replaces the discrete probabilities $P(X = j)$.

See the appendix to this chapter for further details.

For a normal distribution with parameters μ and σ, the integral on the right-hand side of formula (2.2) evaluates to the parameter μ contained in the equation of the normal curve. So the mean of the distribution is equal to the parameter μ. (You might have expected this result: a normal curve is symmetrical about the value $x = \mu$, so the mean is μ.) So the use of the symbol μ for the parameter of a normal distribution reflects the role of the parameter in the model.

In general, given a sample of data, the sample mean $\bar{x}$ is used to estimate μ, the mean of the population from which the sample was drawn. So if we have a sample from a population and we wish to choose a normal distribution to model the variation in the population, then we shall use the sample mean $\bar{x}$ for the parameter μ in the model. For the heights of Cambridge men, the sample mean is approximately 68.9 inches, so we might well choose a normal distribution with parameter $\mu = 68.9$ to model the variation in the heights of all Cambridge men.

2.2 The standard deviation

There are several measures of spread which may be used to summarise the variability in a sample of data. If you have studied MU120, then you will already be familiar with the range, the interquartile range, the relative spread and the standard deviation as different possible measures of spread. As indicated at the beginning of this section, the parameter σ of a normal distribution is the standard deviation of the distribution, so in this subsection we shall restrict our attention to this particular measure of spread. We shall discuss how it is defined and calculated, both for a distribution and for a sample of data.

The population standard deviation

Consider first a population with mean μ. The population standard deviation is a measure of how widely the values in the population are spread about the mean μ; it is defined in terms of the *deviations* of values from μ. The **deviation** of a value x from the mean μ is $x - \mu$; so the squared deviation is written $(x - \mu)^2$. The population standard deviation is defined most easily in terms of a related measure called the population variance. The **population variance** is defined to be the mean squared deviation from the mean: that is, it is the average (i.e. the mean) value of $(x - \mu)^2$ for the whole population.

The word 'mean' is used twice in this phrase in two slightly different roles: deviations from the mean (seen as a place), and then the mean of these deviations, to be calculated.

You have just seen that, for a continuous model, the average or mean of the values x in a population is found by integrating the product $xf(x)$ over all possible values x, where f is the probability density function of the distribution which describes the variation in the population. The mean of the squared deviations $(x - \mu)^2$ is found in a similar way, but this time we integrate the product $(x - \mu)^2 f(x)$ over all possible values of x:

$$\text{population variance} = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx. \tag{2.3}$$

The **population standard deviation** is defined to be the square root of the population variance.

For a normal distribution with parameters μ and σ, with probability density function given by formula (1.1), the integral on the right-hand side of formula (2.3) evaluates simply to σ^2. So since the standard deviation is the square root of the variance, the standard deviation of the distribution is equal to σ, as was stated earlier. As already mentioned, the symbol σ is used generally to denote the standard deviation of a population or of a probability distribution. So the use of the symbol σ for the second parameter of a normal distribution also reflects the role of the parameter in the model.

See the appendix

The sample standard deviation

Now suppose that a sample $x_1, x_2, \ldots, x_n$ of size n has been obtained from a population with unknown mean μ and unknown standard deviation σ. You have seen that the sample mean $\bar{x}$ can be used to estimate the population mean μ. Correspondingly, the *sample standard deviation*, which is denoted by s, can be used to estimate the population standard deviation. But how is the sample standard deviation defined?

Since the population variance is defined to be the mean squared deviation from the population mean μ, by analogy we could define the sample variance to be the mean squared deviation from the sample mean $\bar{x}$. Replacing μ by $\bar{x}$ and dividing the sum of the squared deviations by n to find the mean squared deviation, this would suggest the formula

$$\text{variance} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Hence, taking the square root,

$$\text{standard deviation} = \sqrt{\frac{1}{n}\sum}$$

This is the formula for the standard deviation that was introduced in MU120. However, in practice, this formula tends to underestimate the population standard deviation σ. If we are fitting a normal model to some data, then we would not want to choose a model which underestimates the variability in the population from which the data were drawn. So we would not want to use precisely this formula to choose the value of the parameter σ. It can be shown, though we shall not try to justify it, that this tendency to underestimate the population standard deviation σ can be corrected by using the divisor $n - 1$ rather than n. (When n is large, the difference between the values obtained using n and $n - 1$ is negligible.) This results in the following definitions.

The proof of this result requires ideas and techniques which are beyond the scope of this course.

The **sample variance**, denoted by s^2, is defined to be

$$s^2 = \frac{1}{\quad}\sum$$

and the **sample standard deviation**, denoted by s, is given by

$$s = \sqrt{\frac{1}{\quad}\sum(x - \bar{x})^2} \qquad (2.4)$$

The formula for the sample standard deviation looks complicated, but the idea behind it is straightforward and, in practice, your calculator or your computer will do the actual calculations involved. Most calculators provide two different versions of the standard deviation. One popular brand has two options labelled σ_{n-1} and σ_n; another has two options labelled Sx and σx. These options correspond to the two different formulas: in each case, the first uses the divisor $n - 1$ and the second the divisor n. If your calculator has options for the two different versions of the standard deviation, then make sure you know which option is which. Two exercises are included at the end of this section so that you can check that you can use your calculator to find the sample mean and sample standard deviation (that is, with divisor $n - 1$) of a data set.

Both formulas for the standard deviation are in common use, but they are used in different circumstances. It is usual to use the sample standard deviation s with divisor $n - 1$ as defined above when estimating a population standard deviation σ. So, when a normal distribution is used to model the variation in a population, the sample standard deviation s with divisor $n - 1$ is used for the parameter σ in the model. For the heights of Cambridge men, the sample standard deviation is approximately 2.57 inches, so we might well choose a normal distribution with parameter $\sigma = 2.57$ to model the variation in the heights of all Cambridge men.

In this course, we shall always use the divisor $n - 1$ when calculating the standard deviation of a sample of data. The statistics software OUStats also uses the divisor $n - 1$.

Summary of Section 2

In general, given a sample of data from a population, the sample mean $\bar{x}$ is used to estimate the population mean μ, and the sample standard deviation s is used to estimate the population standard deviation σ. This is summarised in the table below.

Sample statistic		Population parameter
$\bar{x}$	is used to estimate	μ
s	is used to estimate	σ

The sample statistics $\bar{x}$ and s are defined as follows.

For a sample of n observations $x_1, x_2, \ldots, x_n$, the sample mean $\bar{x}$ is given by

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the sample standard deviation s is given by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

More generally, if $x_1, x_2, \ldots, x_k$ are the different values observed in a sample of n observations, and $f_1, f_2, \ldots, f_k$ are their respective frequencies (so $f_1 + f_2 + \cdots + f_k = n$), then the sample mean $\bar{x}$ is given by

$$\bar{x} = \frac{1}{n}(x_1 f_1 + x_2 f_2 + \cdots + x_k f_k) = \frac{1}{n}\sum_{i=1}^{k} x_i f_i$$

and the sample standard deviation s is given by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}$$

When choosing a normal distribution to model the variation in a population, the values of two parameters have to be specified. These parameters are the mean and standard deviation of the normal distribution. When fitting a normal model to data, the values of the sample mean $\bar{x}$ and the sample standard deviation s are used for the parameters of the model.

Exercises for Section 2

It is not always convenient to use a computer to calculate statistics such as the sample mean and the sample standard deviation; and it may not be worthwhile switching on a computer if the data set is fairly small. So it is important for you to be able to perform these calculations on your calculator. If your calculator has options for the sample mean and the sample standard deviation, then make sure that you can use them to carry out the following exercises. If your calculator does not have these statistical facilities, then do the calculations by completing a table with columns headed x_i, $x_i - \bar{x}$, $(x_i - \bar{x})^2$. The sample mean $\bar{x}$ can be found by summing all the numbers in the first column and dividing by n. To find the sample standard deviation, add up the numbers in the third column to find $\sum(x_i - \bar{x})^2$, divide by $n - 1$, and finally find the square root.

Exercise 2.1 How much mail?

The number of items of mail delivered on a Monday morning to each of five houses chosen at random from those in a large estate were as follows.

2 7 3 1 2

(a) Find the sample mean and the sample standard deviation.

(b) If there are 3000 houses on the estate, what is your estimate of the total number of items of mail delivered to the estate on that morning?

Exercise 2.2 Earnings of mechanical engineers

The gross weekly earnings (in pounds) in 1995 of a sample of six mechanical engineers were as follows.

310 635 464 520 381 732

Find the sample mean and the sample standard deviation.

These are invented data, but they are based on statistics published in the 1995 New Earnings Survey.

3 Fitting a normal model

To study this section, you will need access to your computer, together with the statistics software and Computer Book D, and an audio tape player together with Audio Tape 3.

In Section 1, the normal distribution was introduced as a model for the variation in various phenomena, including the heights of Cambridge men, the chest measurements of Scottish soldiers and the lengths of cuckoo eggs. You saw that a particular normal model is chosen by fixing the values of two parameters, μ and σ. In Section 2, we discussed what these parameters represent and how, given a sample of data, their values may be estimated. In this section, the software OUStats will be used to fit normal models to a number of data sets. You will need your computer throughout this section; it has been divided into three subsections to provide suitable break points in case you are unable to complete all the activities in one session.

In Subsection 3.1, the main facilities of OUStats are introduced, using an audio tape session. The data on the heights of Cambridge men are used to illustrate the fitting of a normal model. You will also see how the model can be used to answer questions such as 'what proportion of all Cambridge men in 1902 were between 69.5 and 70.5 inches tall?'.

In Subsection 3.2, you will be asked to explore several data sets and, in each case, to investigate whether a normal distribution provides a good model for the variation observed in the data.

At some stage in your study of this block, you may need to print output from OUStats. Subsection 3.3 contains instructions for printing output from OUStats. It would be a good idea to follow these instructions in order to check that you can use your printer to obtain such output.

Refer to Computer Book D for the work in this section.

for MSP121', during the compu

Summary of Section 3

In this section, you have begun to use OUStats, the data analysis part of the statistics software. You have used OUStats in the following ways.

- ◇ To obtain summary statistics for data.
- ◇ To obtain frequency diagrams for data.
- ◇ To fit a normal curve to data.
- ◇ To calculate areas under a normal curve.
- ◇ To generate random samples from a normal distribution.
- ◇ To print output.

4 Are people getting taller?

To study this section, you will need access to your computer, together with Computer Book D and the statistics software.

In this section, you will have the opportunity to investigate a large data set on the heights of fathers and their sons. The data were collected in the 1890s as part of a larger study into the inheritance of characteristics in humans. The data are of historical interest and have been the subject of much study since they were published.

Refer to Computer Book D for the work in this section.

Summary of Section 4

In this section, you have investigated a large data set on the heights of fathers and sons. You have used many of the facilities of OUStats that were introduced in Section 3. In addition, you have used OUStats to obtain a scatterplot for paired data.

5 Exploring normal distributions

To study this section, you will need access to your computer, together with Computer Book D and the statistics software.

In Section 1, the normal distribution was introduced and some basic properties of normal distributions were described: for example, the position of the 'centre' of a normal curve is given by the parameter μ, which is the mean of the distribution, and the spread of values is governed by the parameter σ, the standard deviation of the distribution. In Sections 3 and 4, you used the statistics software to fit normal models to data and to do calculations for normal distributions. In this section, you are invited to use the software to explore further the properties of normal distributions.

Refer to Computer Book D for the work in this section.

Summary of Section 5

Several general properties of normal distributions have been illustrated in the activities in this section. These and other properties of normal distributions are summarised in the final section of this chapter.

6 Properties of normal distributions

The normal distribution underpins much of the work that you will be doing in Chapters D3 and D4. The purpose of this section is to bring together the main properties of normal distributions that have been discussed and explored in this chapter. We begin by summarising the properties of normal distributions that were discussed in Sections 1 and 2.

The normal distribution was introduced as a model for the variation observed in a number of samples of data, including the heights of Cambridge men, the lengths of cuckoo eggs and the weights of Irish dipper nestlings. In fact, there is not just one normal distribution but a whole family of closely related normal distributions, each being specified by two parameters, most commonly denoted by μ and σ.

The equation of a normal curve is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (-\infty < x < \infty).$$

A typical normal curve is shown in Figure 6.1.

Figure 6.1 A typical normal curve

A normal curve is symmetric about its peak, which occurs at $x = \mu$; so the mean of the distribution is equal to the parameter μ. Normal curves with different means but with the same value of σ are identical in shape but their peaks are located at different positions on the x-axis. Figure 6.2 shows sketches of the three normal curves from Activity 5.1 in Computer Book D; the value of the parameter σ is 1 in each case.

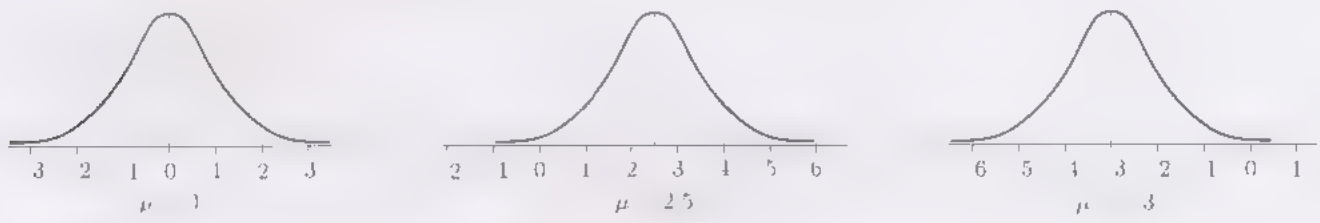
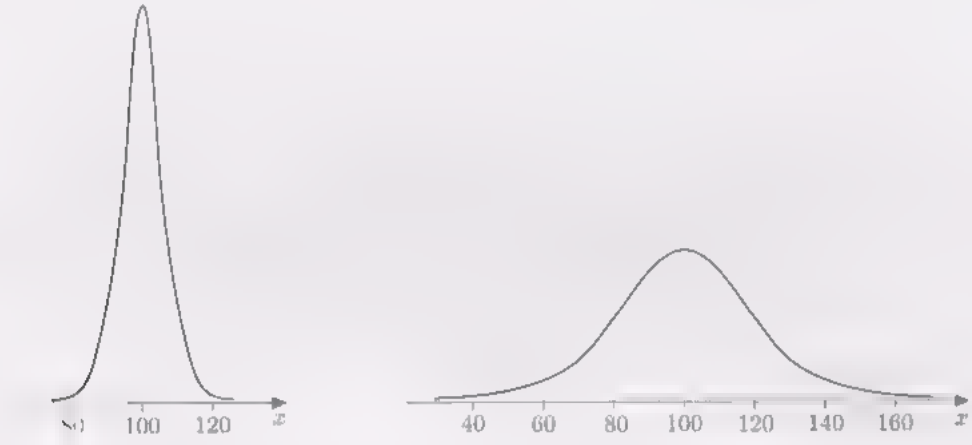


Figure 6.2 Three normal curves each with $\sigma = 1$

The parameter σ of a normal distribution is the standard deviation of the distribution and is a measure of the spread of values in the distribution. Roughly speaking, a small value of σ produces a tall narrow 'bell' and a large value of σ produces a short wide 'bell'. (Although, of course, whether a 'bell' looks short and wide or tall and narrow will depend on the scales used on the axes.) Two normal curves with the same mean but different standard deviations are sketched in Figure 6.3; the same vertical and horizontal scales were used to produce the two sketches.



(a) $\sigma = 6$ (b) $\sigma = 20$

Figure 6.3 Two normal curves (same scales) each with $\mu = 100$

You have seen that an area under a normal curve between $x = a$ and $x = b$ represents the proportion of the population being modelled which have values between a and b. When modelling the heights of Cambridge men in 1902, for example, the area between $x = 69$ and $x = 71$ under the normal curve used to model heights represents the proportion of all Cambridge men who were between 69 and 71 inches tall. This area is the shaded area in Figure 6.4.

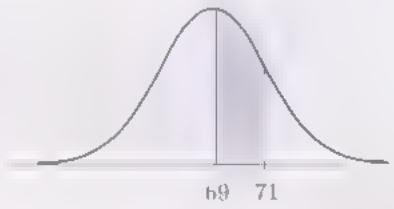


Figure 6.4 An area under a normal curve

An alternative interpretation of this area is as a probability. If a man had been chosen at random from all Cambridge men in 1902, then this area represents the probability that his height would have been between 69 and 71 inches. You found a number of areas under normal curves in Section 4 when investigating the heights of fathers and their sons.

In Section 5, after finding areas under a small number of normal curves, you were invited to make hypotheses about areas for all normal distributions, whatever the values of μ and σ. The results of Activities 5.1 to 5.4 in Computer Book D may be summarised as follows.

Suppose that a normal distribution is used to model the variation in a population. Then according to the model:

- approximately 68.3% of the population are within 1 standard deviation of the mean (that is, between $\mu - \sigma$ and $\mu + \sigma$);
- approximately 95.4% of the population are within 2 standard deviations of the mean (that is, between $\mu - 2\sigma$ and $\mu + 2\sigma$);
- almost all the population – about 99.7% – are within 3 standard deviations of the mean (that is, between $\mu - 3\sigma$ and $\mu + 3\sigma$).

Note that these results hold true whatever the values of the mean μ and the standard deviation σ of the distribution. The results are illustrated in Figure 6.5.

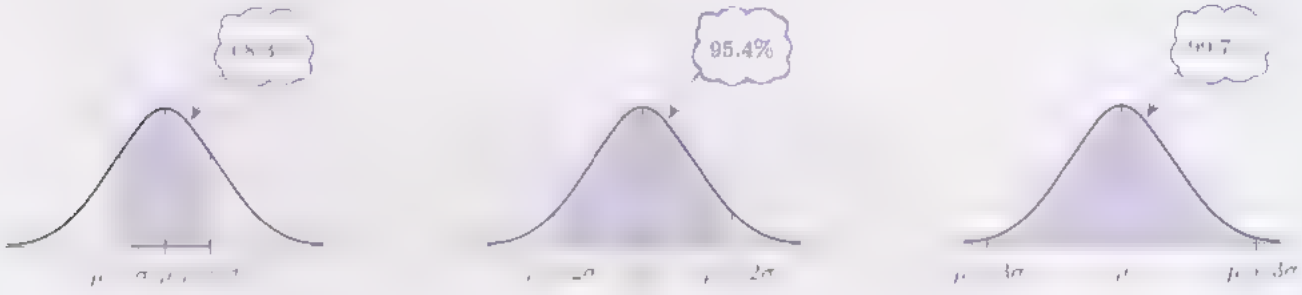


Figure 6.5 Areas under normal curves

Example 6.1 Heights of Cambridge men

The normal distribution used to model the heights of Cambridge men in 1902 has mean $\mu = 68.9$ and standard deviation $\sigma = 2.57$. According to this model, within what range were the heights of almost all Cambridge men – that is, about 99.7% of them? How does this compare with the sample of heights?

Solution

Almost all Cambridge men - about 99.7% of them - were between $68.9 - 3 \times 2.57$ and $68.9 + 3 \times 2.57$ inches tall, that is, between approximately 61.2 and 76.6 inches.

So the model predicts that about three in a thousand men will be either shorter than 61.2 inches or taller than 76.6 inches. As you saw, for the sample of 1000 men from this population, only one man's height was outside the range 61.2 inches to 76.6 inches - no man was less than 61.5 inches tall, one man was recorded as 77 inches tall and no man was taller than this. So 'almost all' the men in the sample were within three standard deviations of the mean height. The normal model reflects quite well the proportion of men in the sample who were unusually short or tall.

Activity 6.1 Lengths of cuckoo eggs

In Activity 3.5 in Computer Book D, a normal distribution with mean $\mu = 22.4$ and standard deviation $\sigma = 1.08$ was fitted to the data on the lengths of cuckoo eggs. According to the model, within what range are the lengths of almost all cuckoo eggs (that is, 99.7% of them)? How does this compare with the sample of lengths?

A solution is given on page 41.

Further similar results for a population modelled by a normal distribution with mean μ and standard deviation σ were illustrated in Activities 5.5 to 5.7 in Computer Book D. These may be summarised as follows:

- approximately 90% of the population are within 1.64 standard deviations of the mean, that is, between $\mu - 1.64\sigma$ and $\mu + 1.64\sigma$;
- approximately 95% of the population are within 1.96 standard deviations of the mean, that is, between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$;
- approximately 99% of the population are within 2.58 standard deviations of the mean, that is, between $\mu - 2.58\sigma$ and $\mu + 2.58\sigma$.

Again, these results hold whatever the values of the mean μ and the standard deviation σ. The results are illustrated in Figure 6.6.

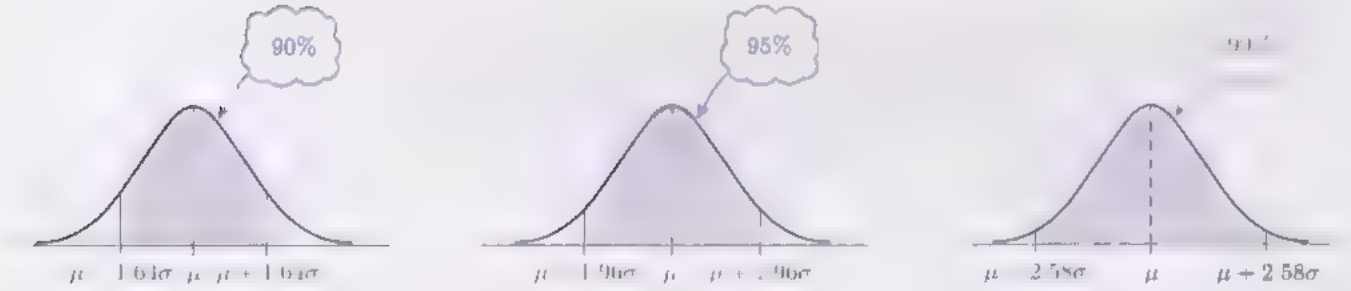


Figure 6.6 Areas under normal curves

These general results are of practical importance and, as you will see, we shall make use of the second result in particular in the next two chapters.

Example 6.2 Heights of Cambridge men

According to the normal distribution used to model the variation in the heights of Cambridge men in 1902, within what ranges were the heights of approximately 90%, 95% and 99% of Cambridge men?

Solution

Applying the general results above to the normal model for the heights of Cambridge men in 1902, we have the following:

- ◇ approximately 90% of Cambridge men were between $68.9 - 1.64 \times 2.57 \simeq 64.7$ and $68.9 + 1.64 \times 2.57 \simeq 73.1$ inches tall;
- ◇ approximately 95% of Cambridge men were between $68.9 - 1.96 \times 2.57 \simeq 63.9$ and $68.9 + 1.96 \times 2.57 \simeq 73.9$ inches tall;
- ◇ approximately 99% of Cambridge men were between $68.9 - 2.58 \times 2.57 \simeq 62.3$ and $68.9 + 2.58 \times 2.57 \simeq 75.5$ inches tall.

Activity 6.2 Lengths of cuckoo eggs revisited

According to the normal model used in Activity 6.1, within what ranges are the lengths of the following percentages of all cuckoo eggs?

(a) 90% (b) 95% (c) 99%

A solution is given on page 42.

All the results about the proportion of values within a given number of standard deviations of the mean are special cases of the following general result.

> If a normal distribution is used to model the variation in a population, then, according to the model, the proportion of the population within k standard deviations of the mean is the same whatever the values of the mean μ and the standard deviation σ.

This result means that the area shaded in Figure 6.7 depends on the value of k, but not on the values of μ and σ. (See the box on page 36 for a suggestion on how you could check this result.)

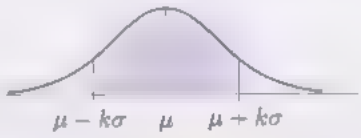


Figure 6.7 k standard deviations from the mean

In particular, for any values of μ and σ, this area is equal to the area between $-k$ and k under the curve of a normal distribution with mean zero and standard deviation one ($\mu = 0$ and $\sigma = 1$). This is illustrated in Figure 6.8.

Figure 6.8 Equal areas

The normal distribution with mean equal to zero and standard deviation equal to one is called the **standard normal distribution**. A sketch of the standard normal curve is shown in Figure 6.9.

Figure 6.9 The standard normal distribution

Tables giving numerical areas under the standard normal curve are included in many statistics textbooks and in books of statistical tables. Since the area within k standard deviations of the mean is the same for all normal distributions, these tables can be used to find areas under *any* normal curve. Even though statistics software packages are available to do calculations for normal distributions, printed tables are still used when it is not convenient, or not worthwhile, to switch on a computer. If you study statistics in the future, you will almost certainly learn how to use such tables. However, in MST121, the computer will be used for all calculations involving normal distributions.

Checking a result

If you are interested, you could check the result concerning the proportion of values that are within k standard deviations of the mean in a normal distribution by finding the area under a normal curve between $\mu - k\sigma$ and $\mu + k\sigma$. This area is given by the integral

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{\mu-k\sigma}^{\mu+k\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx$$

You can use Mathcad to evaluate this integral. You should find that the result given by Mathcad contains k but does not involve μ and σ.

Summary of Section 6

In this section, some properties of normal distributions have been summarised. The main result that will be used in the next two chapters is as follows.

> For a population modelled by a normal distribution, the proportion of the population within k standard deviations of the mean (that is, between $\mu - k\sigma$ and $\mu + k\sigma$) is the same whatever the values of μ and σ.

In particular, approximately 95% of the population are within 1.96 standard deviations of the mean, that is, between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$.

Exercise for Section 6

Exercise 6.1 Sons' heights

In Section 1 you explored Pearson's data on the heights of father–son pairs. For fathers who were 67 inches tall, the mean height of their sons was 68.0 inches and the standard deviation was 2.21 inches. Suppose that a normal distribution with parameters $\mu = 68.0$ and $\sigma = 2.21$ is used to model the heights in inches of the sons of fathers who are 67 inches tall.

You do not need MINITAB for this exercise.

(a) Calculate a range of values within which, according to the model, the heights of 95% of sons of 67-inch-tall fathers lie.

(b) A 67-inch-tall man has a son who is 63 inches tall. How many standard deviations below the mean height for sons of 67-inch-tall fathers is this son, according to the model? Another 67-inch-tall man has a son who is 72 inches tall. How many standard deviations above the mean height for sons of 67-inch-tall fathers is this son? Would you regard either of these sons as unusually short or tall for the sons of 67-inch-tall fathers?

Summary of Chapter D2

In this chapter, the idea of using a curve to model the variation observed in data has been introduced. One particular model – the normal distribution – has been discussed in some detail. But you should remember that this is not the only possible model. In Activity 1.2, you saw examples where differently-shaped curves are needed. There are other distributions which are widely used for modelling variation, and if you study statistics in the future you will meet some of them. However, the normal distribution is of great practical importance, as you will see in the next chapter.

You have also seen that sample statistics, such as the sample mean $\bar{x}$ and the sample standard deviation s, may be used to estimate population parameters, such as the population mean μ and the population standard deviation σ. A large part of your work has involved learning to use the data analysis software OUStats.

In the next two chapters, you will be applying the idea of using sample statistics to estimate population parameters; you will also be using some of the properties of the normal distribution discussed in this chapter. Additionally, some further features of OUStats will be introduced.

Learning outcomes

You have been working towards the following learning outcomes.

Terms to know and use

Mode, unimodal and bimodal, skewed, left-skew and right-skew, probability density function, normal curve, normal distribution, sample mean and sample standard deviation, population mean and population standard deviation, sample statistics and population parameters.

Symbols and notation to know and use

The notation $\bar{x}$ for the sample mean and s for the sample standard deviation.

The symbol μ for the mean of a probability distribution or for a population mean.

The symbol σ for the standard deviation of a probability distribution or for a population standard deviation.

Mathematical skills

- Given a frequency diagram for a data set, sketch a possible model for the variation in the data.
- Interpret areas under the graph of a probability density function.
- Calculate the mean and standard deviation of a small data set using a calculator.
- Investigate the fit of a normal model to a given data set using OUStats.
- Given the parameters of a normal distribution, calculate a range of values within which 90%, 95% or 99% of values lie.

Features of OUStats to use

- ◇ Open and close an existing data file, and obtain information about the data; obtain summary statistics for the data in a file.
- ◇ Obtain a frequency diagram for a sample of data, and fit a normal curve to the data.
- ◇ Find areas under a normal curve.
- ◇ Generate random samples from a normal distribution.
- ◇ Print output.
- ◇ Obtain a scatterplot for paired data.

Ideas to be aware of

- ◇ The uncertainty about the value of a continuous variable can be represented by a curve, that is, by a probability density function.
- ◇ An area under the graph of a probability density function (such as that of the normal distribution) can be interpreted as a probability or as a proportion.
- ◇ The sample statistics $\bar{x}$ and s may be used to estimate the values of the population parameters μ and σ.
- ◇ For a population modelled by a normal distribution, the proportion of the population within k standard deviations of the mean (that is, between $\mu - k\sigma$ and $\mu + k\sigma$) is the same whatever the values of μ and σ. In particular, approximately 95% of the population are within 1.96 standard deviations of the mean.
- ◇ As a rough guide, it is reasonable to quote sample statistics to one significant figure more than is given in the data used to calculate them.

Appendix: Integrals with infinite limits and the normal distribution

In Section 1, the area between a normal curve and the x-axis is defined as an integral with infinite limits. And in Section 2, the mean and variance of a continuous distribution with probability density function f are defined as integrals with infinite limits. This appendix explains briefly one approach to defining an integral with infinite limits, and looks again at the integrals introduced in Sections 1 and 2

1. Integrals with infinite limits

Suppose that, for any number $N > 0$, the definite integral

$$\int_{\quad}^{\quad} f(x)\,dx$$

can be calculated. If the value of this integral approaches a limiting value as N becomes larger and larger, then the expression

$$\int_{-\infty}^{\infty} f(x)\,dx$$

is defined to be equal to this limiting value; that is,

$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{N \to \infty} \int_{-N}^{N} f(x)\,dx.$$

2. The total area under a normal curve

For a normal distribution with probability density function f given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (-\infty < x < \infty),$$

the total area between the normal curve and the x-axis is given by

$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{\quad} \int_{\quad}^{\quad} f(x)\,dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{\quad}^{\quad} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx$$

3. The mean and standard deviation of a normal distribution

In Subsection 2.1, it was stated that the mean of a normal distribution is given by the value of the parameter μ. This result can be deduced using the symmetry of the normal curve, but the mean is defined by the integral

$$\int_{-\infty}^{\infty} x f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx.$$

In Subsection 2.2, it was stated that the standard deviation of a normal distribution is given by the parameter σ. The variance is defined by the integral

$$\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x \quad \mu)^2 \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx,$$

and the standard deviation is the square root of the variance.

Solutions to Activities

Solution 1.2

For the earthquake data in Figure 1.4(a), short times are the most common, and longer times occur less often. The frequency diagram has a long right tail. A sketch of a possible model for times between major earthquakes is given in Figure S 1.

Figure S.1 A possible model for times between earthquakes

The frequency diagram for the wing lengths of first-year meadow pipits has a peak at about 81 mm, and there is a suggestion that there might be a second smaller peak at about 76 or 77 mm. A possible model for wing length is sketched in Figure S.2.

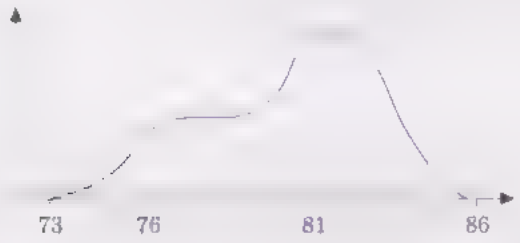


Figure S.2 A possible model for wing lengths of meadow pipits

The frequency diagram for the durations of eruptions of Old Faithful geyser has two peaks or modes, at approximately 1.8 and 4.0 minutes; it is *bimodal*. A sketch of a possible model is shown in Figure S.3.

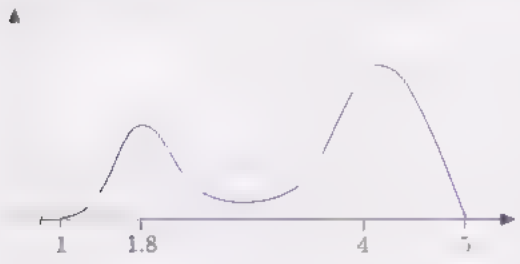


Figure S.3 A possible model for the durations of eruptions of Old Faithful geyser

The frequency diagram for the women's weekly earnings data has a single peak at about £200. It has a longer right tail than left tail. A possible model for the variation in weekly earnings is shown in Figure S.4

Figure S.4 A possible model for women's weekly earnings in April 1994

Solution 1.4

(a) The shaded area represents the proportion of all Cambridge men in 1902 who were between 69 and 71 inches tall. Alternatively, it represents the probability that a man selected at random from all Cambridge men would have been between 69 and 71 inches tall.

(b) The shaded area represents the proportion of all Cambridge men in 1902 who were under 65 inches tall. Alternatively, it represents the probability that a man selected at random from all Cambridge men would have been less than 65 inches tal..

(c) The shaded area represents the proportion of all Cambridge men in 1902 who were over six feet tall. Alternatively, it represents the probability that a man selected at random from al. Cambridge men would have been over six feet tall

Solution 6.1

According to the model, the lengths of about 99.7% of all cuckoo eggs should be between $\mu - 3\sigma$ and $\mu + 3\sigma$; that is, between $22.4 - 3 \times 1.08 \simeq 19.2$ mm and $22.4 + 3 \times 1.08 \simeq 25.6$ mm

For the sample of 243 eggs, no egg was as long as 25.6 mm and no egg was shorter than 18.75 mm. One was recorded as 19 mm long (between 18.75 mm and 19.25 mm), so the length of at most one of the 243 eggs was outside the range 19.2 mm to 25.6 mm. So 'almost all' the eggs were within three standard deviations of the mean.

Solution 6.2

(a) The lengths of 90% of eggs should lie between
$22.4 - 1.64 \times 1.08 \simeq 20.6\,\text{mm}$ and
$22.4 + 1\ 64 \times 1.08 \simeq 24\ 2\,\text{mm}$

(b) The lengths of 95% of eggs should lie between
$22.4 - 1\ 96 \times 1.08 \simeq 20.3\,\text{mm}$ and
$22.4 + 1\ 96 \times 1.08 \simeq 24.5\,\text{mm}$.

(c) The lengths of 99% of eggs should lie between
$22.4 - 2.58 \times 1.08 \sim 19.6\,\text{mm}$ and
$22.4 + 2\ 58 \times 1.08 \sim 25.2\,\text{mm}$.

Solutions to Exercises

Solution 2.1

(a) The sample statistics are

$$\bar{x} = 3, \quad s \simeq 2.345\,207\,88 \simeq 2.3.$$

The details of the calculations are given below for you to check your working if your calculator does not have statistical facilities for the mean and standard deviation.

x	$x - \bar{x}$	$(x - \bar{x})^2$
2	−1	1
7	4	16
3	0	0
1	−2	4
2	−1	1
15		22

From the first column of the table, we can calculate

$$\bar{x} = 15/5 = 3.$$

Then the values in the second and third columns can be calculated, to give

$$s = \sqrt{22/4} \simeq 2.3.$$

(b) The sample mean $\bar{x}$ may be used to estimate the mean number of items of mail delivered to each house on the estate: 3. So an estimate of the total number of items of mail delivered to the estate on that morning is $3 \times 3000 = 9000$.

Solution 2.2

The sample statistics are

$$\bar{x} = 507, \quad s \simeq 157.297\,171 \simeq 157.3.$$

The details of the calculations are given below.

x	$x - \bar{x}$	$(x - \bar{x})^2$
310	−197	38 809
635	128	16 384
464	−43	1 849
520	13	169
381	−126	15 876
732	225	50 625
3042		123 712

From the first column of the table, we can calculate

$$\bar{x} = 3042/6 = 507.$$

Then the values in the second and third columns can be calculated, to give

$$s = \sqrt{123\,712/5} \simeq 157.297\,171 \simeq 157.3.$$

So the mean gross weekly earnings in 1995 of the sample of six mechanical engineers was £507, and the sample standard deviation of gross weekly earnings was approximately £157.30.

Solution 6.1

(a) According to the model, the heights of 95% of sons will be within 1.96 standard deviations of the mean, that is, between
$68.0 - 1.96 \times 2.21 \simeq 63.7$ inches and
$68.0 + 1.96 \times 2.21 \simeq 72.3$ inches.

(b) At 63 inches tall, the first son is 5 inches shorter than the mean height (68 inches) of sons of 67-inch-tall fathers, that is, $5/2.21 \simeq 2.26$ standard deviations below the mean.

At 72 inches tall, the second son is 4 inches taller than the mean height of sons of 67-inch-tall fathers, that is, $4/2.21 \simeq 1.81$ standard deviations above the mean.

One way to look at the last question is as follows. Since the heights of 95% of sons are within 1.96 standard deviations of the mean height, and only 5% are 1.96 or more standard deviations from the mean height, we might regard a son whose height is at least 1.96 standard deviations above the mean as unusually tall and a son whose height is at least 1.96 standard deviations below the mean as unusually short. On this basis, the first son is unusually short; but we would not consider the second son to be unusually tall, as his height is less than 1.96 standard deviations above the mean.

Index

The Open University
ISBN 0 7492 6662 7